

PROBABILISTIC CROSS-IDENTIFICATION OF COSMIC EVENTS

TAMÁS BUDAVÁRI

Dept. of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218; budavari@jhu.edu
Draft version January 13, 2013

ABSTRACT

We discuss a novel approach to identifying cosmic events in separate and independent observations. In our focus are the true events, such as supernova explosions, that happen once, hence, whose measurements are not repeatable. Their classification and analysis have to make the best use of all the available data. Bayesian hypothesis testing is used to associate streams of events in space and time. Probabilities are assigned to the matches by studying their rates of occurrence. A case study of Type Ia supernovae illustrates how to use lightcurves in the cross-identification process. Constraints from realistic lightcurves happen to be well-approximated by Gaussians in time, which makes the matching process very efficient. Model-dependent associations are computationally more demanding but can further boost our confidence.

Subject headings: methods: statistical — methods: data analysis — surveys — time — supernovae: general

1. INTRODUCTION

Advancements in detector technology have recently opened up new opportunities for time-domain astronomy. The next-generation surveys will observe most of the sky not once but many times over, which enables for systematic studies of variable and transient sources. The first such programs are already online. One example is the Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009) that uses three dedicated telescopes, and has reported over two thousand new transients to date. Another is the Palomar Transient Factory (PTF; Rau et al. 2009) that has discovered close to 800 supernovae. The Pan-STARRS telescope that started survey operation in 2010, will see thousands of supernovae a month, not to mention the SkyMapper project (Keller et al. 2007) or the Large Synoptic Survey Telescope (LSST) that aims to make the “greatest movie of the Universe” by covering most of the sky in every few days.

Modern surveys implement automated discovery in their pipelines. They provide live streams of events in both machine and human readable formats to facilitate rapid follow-up observations and real-time analysis. An interoperable representation is the VOEvent format defined by the International Virtual Observatory Alliance (IVOA, see www.ivoa.net). Capitalizing on this standard, the VOEventNet.org project promises to deliver alerts to interested subscribers within minutes of discovery. Another important reason for standardized live streams is the joint analysis of the events. Users of the SkyAlert.org site can browse for discoveries in a growing selection of data streams. The project will automatically associate any events within 10 degrees and less than 1 hour apart (Roy Williams, private communication).

One outstanding issue is the proper statistical foundation of event aggregation. The methods developed for cross-identification of catalogs are not directly applicable. While positional information is clearly needed for transient events, too, one cannot statistically interpret

the spatial thresholds. To make a probabilistic statement one has to consider the density but the host objects might not be visible, only the explosions. In fact, events really are more abstract entities in spacetime, whose density we have to study in 3+1 dimensions. Rigorous statistical assessments are most important in the presence of crowding: large density compared to the uncertainties. Spatially the optical observations are typically much more accurate than the X-ray, radio, let alone gamma or gravitational wave detections. In time, this is especially relevant considering that the frequency of events depends on the cutoff parameter of the detection. By lowering the threshold we can study fainter and more distant events that are although low significance in any one survey, could be interesting when seen in many. In this paper, our goal is to derive the probabilistic matching formalism for transient events, which enables the automated association of detections in space and time.

In Section 2 we use Bayesian hypothesis testing to evaluate the quality of candidate associations. Section 3 describes why it is *not* impossible after all to derive probabilities for these matches. In Section 4 we analyze simulated SN Ia lightcurves to derive time constraints, and show an accurate approximation that provides analytic results. Section 5 concludes our study. Throughout the paper, the capital P and L symbols represent probabilities and likelihoods, respectively, and the lower-case p letters are probability densities.

2. MATCHING EVENTS

Many things happen in the sky. Asteroids move, stars pulsate, supernovae and gamma-ray bursts explode all over the place. Such astronomical events are typically discovered in photometric images based on the change in the apparent brightness between different epochs. The imaging cadence is strategically planned to meet the scientific goals. The epochs are optimized for the characteristic time-scale of the intended targets, which might be weeks, days, or just tens of minutes. Looking at repeated observations, we can spot the significant changes, e.g., using a 5σ threshold. At that point, it is possible

go back to extract the lightcurve of the detected event from the earlier images. Variable sources are revisited several times over the course of a survey, which provides ample amount of data. Cosmic events, however, happen just once and we need to make the best use of the observations.

Considering that different photometric pass-bands potentially probe different physical processes, the lightcurves can be quite a bit different. The detections can be delayed from one instrument to another. One thing, however, is common: the start time of the event, which can be estimated in all detections by looking at the changing flux as function of time.

Let us now consider a single candidate association that consists of n reported events, one from each survey i (where $i=1$ to n). They will have measured $D_x = \{x_i\}$ positions in the sky (unit vectors of the directions) and can constrain $D_t = \{t_i\}$ start times with some uncertainties. The time constraint is not the precision of our clock, but rather the ability of our model M of the given survey to estimate the start time t of a particular event. Formally, it is a likelihood $L(\tau|t) \equiv p(t|\tau, M)$ that is a function of the model's τ epoch. For now, let us assume that $L_i(\tau|\cdot)$ is known for every survey. In Section 4 we take a closer look at how to derive such likelihood functions from observed lightcurves; in particular we analyze the Type Ia supernovae in detail.

We ask the question whether the detections in the given tuple are from the same event or not using Bayesian hypothesis testing. We wish to compare two hypotheses. The first one, call it H , claims that all detections are of the same cosmic event, hence they have a common epoch (τ) and a common true position (both unknown to us.) The second hypothesis is the former's complement K that allows for any one of the detections to be from a separate event, which we model using a set of parameters, e.g., the $\{\tau_i\}$ epochs. The Bayes factor makes the comparison, which is the ratio of the probabilities of the measurements in the two hypotheses. For independent measurements, we can calculate separate Bayes factors for the position and time measurements, and the combination of the two is just their product,

$$B = \frac{p(D|H)}{p(D|K)} = \frac{p(D_t|H)}{p(D_t|K)} \frac{p(D_x|H)}{p(D_x|K)} = B_t B_x \quad (1)$$

The B_x term is the same as for the static objects as discussed in detail by Budavári & Szalay (2008). The numerator and the denominator of B_t are integrals of the known likelihood functions multiplied by the prior,

$$p(D_t|H) = \int d\tau \, p(\tau|H) \prod_i^n L_i(\tau|t_i) \quad (2)$$

$$p(D_t|K) = \prod_i^n \int d\tau_i \, p(\tau_i|K) L_i(\tau_i|t_i) \quad (3)$$

The choice of the prior naturally affects the Bayes factor, hence it requires some more consideration.

Common sense dictates that the prior on the epoch τ should be flat as the event can occur at any time with the same probability density. The calculation of the Bayes factor, however, requires a proper prior, whose integral is unity. Considering that all relevant measurements and

the non-zero support of their uncertainties can be constrained to a finite interval without any practical limitations, we can introduce a flat prior on an arbitrary but wide enough $[T, T+\Delta]$ interval, and define

$$p(\tau|\cdot) = \begin{cases} \Delta^{-1} & \text{if } \tau \in [T, T+\Delta] \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Using this flat prior, the likelihoods of the hypotheses become

$$p(D_t|H) = \Delta^{-1} \int d\tau \prod_i^n L_i(\tau|t_i) \quad (5)$$

$$p(D_t|K) = \Delta^{-n} \prod_i^n \int d\tau_i L_i(\tau_i|t_i) \quad (6)$$

where we omitted the integration limits as they are irrelevant for any large enough interval, and arrive at the dependence of

$$B_t \propto \Delta^{n-1} \quad (7)$$

This means that the wider we make the interval, the larger the Bayes factor becomes, which seems to artificially go against our expectation for an objective quality measure of the associations.

3. STREAMS OF EVENTS

Cosmic events should not be looked at in isolation. A given telescope provides an entire *stream* of events, which is typically published online in an automated fashion. Our goal is to merge several independent streams of many surveys. The resulting combined stream will carry more information on the individual events, which would help with their scientific analysis. Along with our stream of associations we want to include the probabilities.

The rate of events varies from survey to survey. For a given time interval the number of occurrences can be quite different. Indeed this is a key piece of information that enters the calculations. By definition the posterior probability of an association is computed from the Bayes factor and the prior of the hypothesis. If we assume that the prior probability is uniform, i.e., it takes the same value independent of the position in the sky, etc., we can write it as the ratio

$$P_0 = \frac{N_\star}{\prod_i N_i} \quad (8)$$

just like in the static case. Here N_i is the number of events in the i th survey in a given time interval Δ , and N_\star is the number of events that appear in all surveys, i.e., the intersection of their selection functions. In our case it is only natural to use ν frequencies instead, and substitute $N = \nu\Delta$ into the prior,

$$P_0 = \frac{\nu_\star}{\prod_i \nu_i} \Delta^{1-n} \quad (9)$$

The probability shrinks with increasing time intervals, i.e., it is less probable to randomly select the same event from larger sets. In practice, the number of events is always large, otherwise their frequency could not be measured. More formally, Δ can be arbitrarily wide to yield large N_i counts and a small prior. For vanishing priors,

the posterior depends only on the product of the P_0 prior and the Bayes factor

$$P \simeq \frac{P_0 B}{1 + P_0 B} \quad (10)$$

hence the dependence on Δ actually cancels out in the product of

$$P_0 B_t = \left(\frac{\nu_\star}{\prod \nu_i} \right) \left[\frac{\int d\tau \prod L_i(\tau|t_i)}{\prod \int d\tau_i L_i(\tau_i|t_i)} \right] = \pi_0 \beta_t \quad (11)$$

The second term of the product, β_t , can be directly calculated from the data. The value of ν_\star in π_0 can also be estimated over time from the ensemble statistics of the streams. We recall the self-consistency argument of Budavári & Szalay (2008): the sum of the priors over the all possible $\prod N_i$ combinations, which is N_\star in total by definition, is also equal to the sum of the posteriors. This equality is an equation for N_\star , or ν_\star in this case, that in turn provides well-determined prior and posterior probabilities. In summary, the posterior is well-defined and calculated as

$$P \simeq \frac{\pi_0 \beta_t B_x}{1 + \pi_0 \beta_t B_x} \quad (12)$$

To select candidate associations early on, when not enough statistics are available to reliably solve for ν_\star , one can use the upper bound of $\nu_\star^{\max} = \min\{\nu_i\}$ to overestimate the posteriors. Similarly if ν_i are uncertain, one can use conservative limits instead. By safely overestimating the prior, we end up with somewhat larger posteriors than the actuals, thus a fix probability cut yields slightly more candidates that we can later prune. This will not overshoot by too much because large posteriors are not particularly sensitive to small variations in the prior (Heinis et al. 2009).

4. CONSTRAINTS FROM LIGHTCURVES

One cannot directly measure the epoch of an event, only the fluxes as a function of time. If we can model the change, these lightcurves can provide the time constraints. For example, a Type Ia supernova is described by its absolute magnitude M , redshift z and epoch τ . For any given set of such parameters we can derive the fluxes at the times of pre-scheduled observations, and compare them to reality. The likelihood function is given by the deviation of the photometric measurements f and the simulated fluxes $\bar{f}(\tau, M, z)$ as

$$L(\tau, M, z|f) = N(f|\bar{f}(\tau, M, z), \Sigma) \quad (13)$$

where Σ represents the photometric errors in the normal distribution N . By integrating over M and z with their proper priors, we can derive the τ dependence needed for β_t .

We use a Gaussian luminosity function (Dahlen et al. 2004) with $M_\star = -19$ and $\sigma_M = 0.5$ for the prior, and simulate a typical M_\star SN Ia using the LSST cadence (Ivezic et al. 2008). In this cadence, pairs of observations separated by an hour are repeated every 3 days. Throughout this model we use precision LSST lightcurve templates from Peter Nugent (via Andy Connolly, private communication) but consistently neglect the K-correction. The error on the g -band photometry is determined from the 5σ limiting magnitude of 25. The

simulated event is set to happen at 0 day. The luminosity function determines the kind of SN Ia events that can occur, and the decline at high redshifts is set by the limiting magnitude of the survey. Together these set the integration domain for our calculations. The peak brightness of the simulated supernova at redshift $z=0.8$ is 24.5 magnitudes in the g band. This distance is approximately at the peak of the redshift distribution.

Figure 1 shows the likelihood as a function of the epoch for a noiseless set of observations. The results are not sensitive to the end date as long as the integration limit is in the far tail of the decay. The measurements (*open circles*) follow a Gaussian surprisingly well. The *solid line* helps to guide the eye: the curve has $\sigma=1$ day, and a mean of 0.1 days. Random realizations of lightcurves with realistic noise yield similar profiles with maxima, whose distribution is consistent with the error. For alerting the transient community, the convention is to have 2 independent observations that pass the detection threshold. For LSST-like cadences, these most likely will happen on the same night 1 hour apart. As part of the notification message, a lightcurve can be published that consists of the measurements up to the last two 5σ observations. The *crosses* show the results for such truncated lightcurves, which also closely follow a Gaussian with $\sigma=3$ days and a 0.5 day mean (*dotted line*). The shorter time baseline and the marginal photometry provide significantly less leverage on the time constraint. We also see more deviation, and a slight asymmetry in the points. The offset in the peak at these faint magnitudes is a property of the lightcurve's shape and the cadence. When the photometric errors are tighter, the shift (and the width) is (are) significantly smaller. In comparison, for a $z=0.5$ supernova (with the same M_\star luminosity) the brighter lightcurve yields a negligible offset and a $\sigma \sim 0.25$ days. More frequent sampling also decreases the offset but it is not practical in real life. A better way is to calibrate the effect and correct for the shifts, if needed.

Using Gaussian likelihoods the integrals in β_t can be calculated analytically. For a 2-way association it becomes

$$\beta_t = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp \left\{ -\frac{(t_1 - t_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right\} \quad (14)$$

This is not a Bayes factor and is not dimensionless. Substituting $\sigma=1$ day and $\Delta t=0$ yields $\beta_t = 1/\sqrt{4\pi} \text{ day}^{-1}$, which is 0.282/day approximately. The numerical integral of the photometric likelihood gives the similar 0.287/day value, which is in good agreement with the analytical estimate. The *left panel* of Figure 2 illustrates β_t as a function of Δt for Gaussians with the aforementioned σ values. The *solid line* shows results from matching with the complete lightcurves and the *dotted line* is for alerts with truncated photometry. The *dashed line* illustrates the case of matching a full lightcurve versus an alert. The *right panel* shows the same results for a case, when the likelihoods are approximated by tophat functions with the same standard deviations. The finite support of these functions results in $\beta_t=0/\text{day}$ at large separations in time, which is a cause for concern if the width of the window is not well-understood. For a $z=0.5$ redshift M_\star SN Ia with $\sigma=0.25$ days the value is 4 times larger, $\beta_t=1.128/\text{day}$, when seen at the same time.

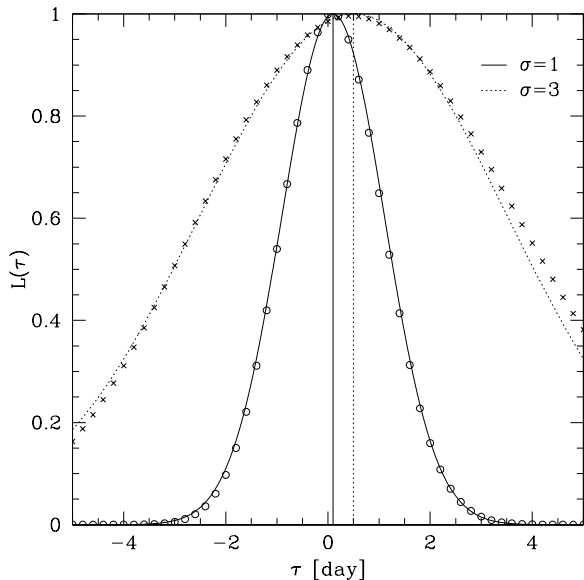


FIG. 1.— The likelihood as a function the epoch is calculated from the lightcurve of a simulated SN Ia, whose observations are in the LSST cadence. The *solid line* is a Gaussian that follows the computed *open circles* well. When the survey detects a 5σ transient, it issues an alert. The *crosses* and the *dotted line* show the same curve for a truncated lightcurve that is only available up until 2 detections are over the threshold. The shorter baseline yields a weaker constraint.

These values of β_t are not large, especially considering that π_0 is typically fairly small, because the time constraint are weak. The frequency of all events from a telescope can be tens of thousands, say up to $\nu \sim 100,000/\text{day}$, when including all transients. Since the true SN Ia rate is expected to be around $\nu_* \sim 1,000/\text{day}$ in LSST, our estimate for the lower bound comes to $\pi_0 \sim 10^{-7}$ days. Nevertheless even for the fainter sources, we have higher $P_t \sim 10^{-8}$ values than in the blind matching of static sources, where $P_0 \sim 10^{-10}$ (Heinis et al. 2009). We can increase our prior by filtering the input streams, however, the filtering might be better done on the associations to find the low significance events in the merged stream. In practice, filtering will be applied both before and after for optimal performance.

For completeness we mention that the general result for the n -way Gaussian case is also calculated analytically. The resulting formula is similarly simple and reads

$$\beta_t = (2\pi)^{\frac{1-n}{2}} \sqrt{\frac{\prod w_i}{\sum w_i}} \exp \left\{ \frac{(\sum w_i t_i)^2}{2 \sum w_i} - \frac{\sum w_i t_i^2}{2} \right\} \quad (15)$$

where $w_i = 1/\sigma_i^2$ are the precision parameters of the t_i epochs.

Photometric measurements can also be directly used to calculate β_t using the multivariate likelihood function in eq.(13). In addition to a common epoch, identical events have the same physical properties as well, e.g., the absolute magnitude and the redshift of the supernova. Our hypotheses H and K remain the same as before, but we parameterize them with not only τ , but also with M and z . To calculation of β_t involves computing the integrals over all three parameters. Of course the numerical integral is substantially more expensive computation-

ally than evaluating an analytic formula but this is the proper use of the model that fully exploits its constraining power. Following the same prescription as before, we generated two lightcurves using the LSST cadences. The only difference was in their phase: the cycle of one simulated survey was shifted by a day from the other. The result for an M_* supernova at $z=0.8$ redshifts is almost 20 times larger than just based on time coincidence, $\beta_t = 5.3/\text{day}$. If we are willing to live with the extra computation and the explicit dependence on a model, this way we can boost the probability, in this case, to almost 1%. Naturally, one would still need spatial constraints for any meaningful match, but could potentially leverage lower accuracy positional measurements.

The explicit modeling of the lightcurve provides an opportunity to simultaneously perform classifications and outlier detection. We can benefit from calculating several Bayes factors or β_t values for comparison. Associations that appear to be reliable matches based on their directions and epochs when using a model independent approximate time window, might be rejected based in a detailed analyses of their SEDs. Accidental noise or new discovery? Follow up observations might help to decide. Beyond the above outlier detection, we can also run several SED models to find out in which class the candidate event might fit best. Naturally there is also nothing to stop us from creating several merged streams using different strategies, which target specific scientific experiments and communities.

5. CONCLUSIONS

We introduced a new statistical method to associate cosmic events based on measured constraints in both time and space. Folding in the time-domain information is important for a number of reasons: (1) If we can automatically cross-correlate streams of events from multiple surveys, we can find fainter ones that are not obvious in any one of the observations but together become significant. This enables us to push these studies to uncharted territories. (2) Constraints on the epoch can boost small positional evidences in case of low accuracies, e.g., gamma-ray bursts. (3) A more fundamental, yet, extremely important practical matter is the assignment of probabilities. While we can clearly calculate Bayes factors based on positional information only, we cannot define the probabilities without the time-domain data. Conceptually the problem is with the prior. It cannot be determined without counting the occurrences of events in a fixed time interval, whose duration is formally arbitrary. We found that the solution is to fold in the Bayes factor in time, which naturally cancels the artifact.

The lightcurve measurements constrain the epoch of a given event. For Type Ia supernovae, we found the likelihood to be very close to a Gaussian that makes the problem analytically tractable. The numerical simulations agree with the fast approximations. The same method also works for streams of events seen in different wavelengths as long as regime is covered by the SED models. To unlock the full potential of the lightcurves one can further exploit the SED modeling. The calculation is extended to require matching events to have the same physical properties. For SN Ia, the numerical integrals show that this strategy has the potential to boost

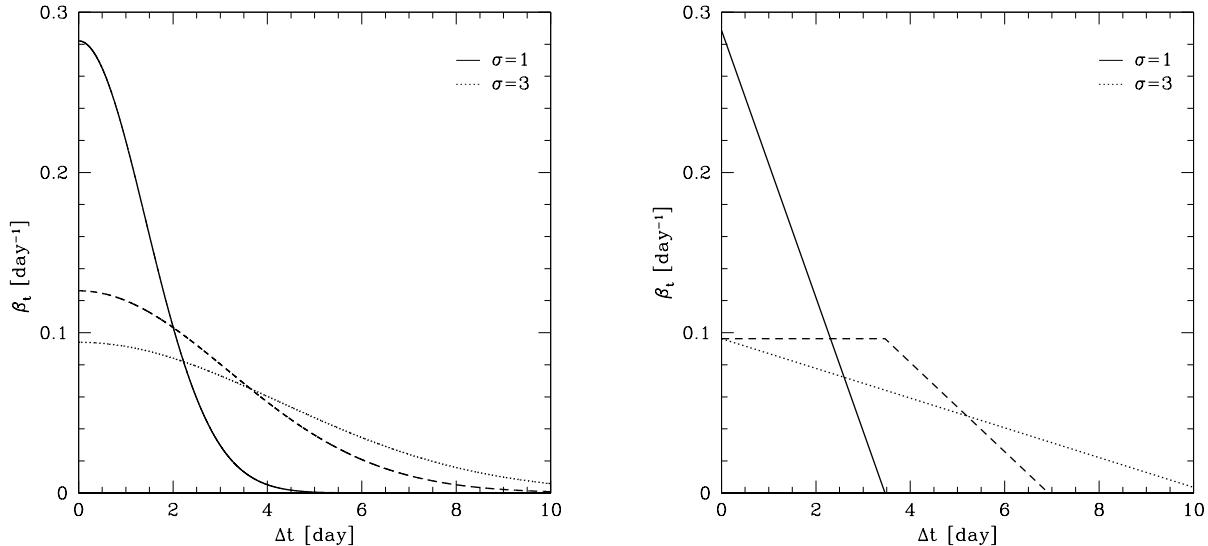


FIG. 2.— When crossmatching streams of events, the β_t parameter is the time-domain analog of the Bayes factor. Its dependence on the time difference is plotted for the 2-way Gaussian case in the *left* panel with uncertainties seen in Figure 1. The line styles are identical to those in Figure 1. The *dashed line* is for matching one against the other. For comparison, the *right* panel shows the same using tophat likelihood approximations with the same standard deviations. The finite support of these functions means that $\beta_t=0/\text{day}$ at large separations in time, which in turn yields 0 posterior probability. The loss of associations, in this case, is a cause for concern, if the width of the time-window is not well-understood.

the time-domain evidence by a factor of 20, which means that one can get away with less accurate spatial measurements.

Although we illustrated the methodology on supernovae, it is clearly applicable to other types of events. Before gamma-ray bursts were established to be extragalactic, Luo et al. (1996) applied Bayesian hypothesis testing to look for repeating gamma-ray bursts. Since then we have seen the host galaxy of many GRBs. Using the time-domain cross-identification we can look for other most subtle events in their hosts from related physical processes. Another example is the detection of gravitational waves. The LIGO Collaboration discusses astrophysically triggered searches (Abbot et al. 2008). Their idea is to lower the detection threshold for a short period of time, when other type of cosmic events occur. This is in fact a very similar problem: merging streams of events from LIGO and other projects with our probabilistic approach could boost their significance and help facilitate

new discoveries.

The general cross-identification problem in astronomy is rather convoluted and ultimately tightly interwoven with the scientific analyses. One obvious example is classification, which in turn would affect the matching prior. The circular nature of the problem is not an issue but one needs to find a consistent solution. The explicit assumptions in the probabilistic approach enables us to incorporate new, different kind of data on top of the epochs and the positions, when available. This is a crucial feature, which will prove vital in studying the time-domain and understanding the eventful Universe.

The author would like to thank Andy Connolly, Alex Szalay, István Csabai, Matthew Graham, Suvi Gezari, Josh Bloom, Joey Richards, Roy Williams and Tom Loredo for several stimulating discussions on various aspects of the topic. This study was supported by the Gordon and Betty Moore foundation via GBMF 554.

REFERENCES

- Abbott, B., et al. 2008, *Class. Quantum Grav.*, 25 114051
 Budavári, T., & Szalay, A. S., 2008, *ApJ*, 679, 301
 Dahlen, T., et al. 2004, *ApJ*, 613, 189
 Drake, A. J., et al., 2009, *ApJ*, 696, 870
 Heinis, S., Budavári, T., & Szalay, A. S. 2009, *ApJ*, 705, 739
 Ivezić, Z., et al. 2008, *Serbian Astronomical Journal*, 176, 1
 Keller, S. C., et al. 2007, *Publications of the Astronomical Society of Australia*, 24, 1
 Luo, S., Loredo, T., & Wasserman, I. 1996, *American Institute of Physics Conference Series*, 384, 477
 Rau, A., et al. 2009, *PASP*, 121, 1334